PROTEINS: Structure, Function, and Genetics 37:582–591 (1999)

Redesigning the Hydrophobic Core of a Model β-Sheet Protein: Destabilizing Traps Through a Threading Approach

Jon M. Sorenson[1] **and Teresa Head-Gordon**[2]*
[1]*Department of Chemistry, University of California, Berkeley, Berkeley, California*
[2]*Division of Physical Biosciences and Life Sciences, Lawrence Berkeley National Laboratory, Berkeley, California*

ABSTRACT **An off-lattice 46-bead model of a small all-β protein has been recently criticized for possessing too many traps and long-lived intermediates compared with the folding energy landscape predicted for real proteins and models using the principle of minimal frustration. Using a novel sequence design approach based on threading for finding beneficial mutations for destabilizing traps, we proposed three new sequences for folding in the β-sheet model. Simulated annealing on these sequences found the global minimum more reliably, indicative of a smoother energy landscape, and simulated thermodynamic variables found evidence for a more cooperative collapse transition, lowering of the collapse temperature, and higher folding temperatures. Folding and unfolding kinetics were acquired by calculating first-passage times, and the new sequences were found to fold significantly faster than the original sequence, with a concomitant lowering of the glass temperature, although none of the sequences have highly stable native structures. The new sequences found here are more representative of real proteins and are good folders in the $T_f > T_g$ sense, and they should prove useful in future studies of the details of transition states and the nature of folding intermediates in the context of simplified folding models. These results show that our sequence design approach using threading can improve models possessing glasslike folding dynamics. Proteins 1999;37:582–591.** **Published 1999 Wiley-Liss, Inc.**†

Key words: protein design; off-lattice models; protein folding; multiple histogram method; multistate kinetics

INTRODUCTION

The use of computational protein models to understand and better characterize protein folding has a long and productive history.[1–7] Recent successful use of such models range in complexity from two-dimensional two-flavor lattice models of prion proteins,[8] to fully atomistic models with explicit solvent molecules for simulating folding and unfolding of small proteins in an aqueous environment.[6,9] Accompanying this growth in computational models, new theoretical perspectives on protein folding, such as the concept of folding funnels[10–12] and the principle of minimal frustration[13] have emerged.

One simplified off-lattice model that has been developed and studied in many contexts is the β-sheet protein proposed by Thirumalai and coworkers.[14–23] The model retains a minimal description of the energetics of the polypeptide backbone and reduces nonlocal interactions to those between beads of three flavors: hydrophobic, hydrophilic, or neutral. Although this simplification of real proteins omits, for example, backbone hydrogen-bonding and side-chain-packing effects, some essential features of the protein-folding problem, such as the role of competing nonlocal interactions and folding to a native state in the presence of an enormous conformational space, are captured.

Recently this model has been criticized for possessing an extremely rough energy landscape with many traps and long-lived intermediates,[20] and a consequently high glass temperature and low folding temperature.[22] The most recent work indicates that the underlying free energy surface for folding is poorly shaped, favoring collapsed states but not sufficiently biasing folding to the native state.[22,23] As real proteins are expected to have folding temperatures higher than their glass temperature ($T_f/T_g \sim 1.3$[24]), the original β-sheet model proposed by Thirumalai and coworkers can be characterized as a poor folder.

The problems with the original model can be traced back to the portions of the sequence which form the hydrophobic core. The dominant element in the native state structure holding the core together are two long stretches of purely hydrophobic beads. The degeneracy of the sequence in this region permits many other possible compact, non-native conformations to exist which possess very similar energetics. In recognition of this problem, some studies have successfully modified the original model by biasing native state contacts to significantly favor formation of the native state.[22,23]

Grant sponsor: Air Force Office of Sponsored Research; Grant number: FQ8671–9601129; Grant sponsor: U.S. Department of Energy (OBER and LDRD); Grant number: DEAC-03–76SFOO098; Grant sponsor: National Energy Research Supercomputer Center; Grant sponsor: National Science Foundation for a Graduate Research Fellowship.

*Correspondence to: Teresa Head-Gordon, Physical Biosciences and Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. E-mail: TLHead-Gordon@lbl.gov

Received 10 May 1999; Accepted 3 August 1999

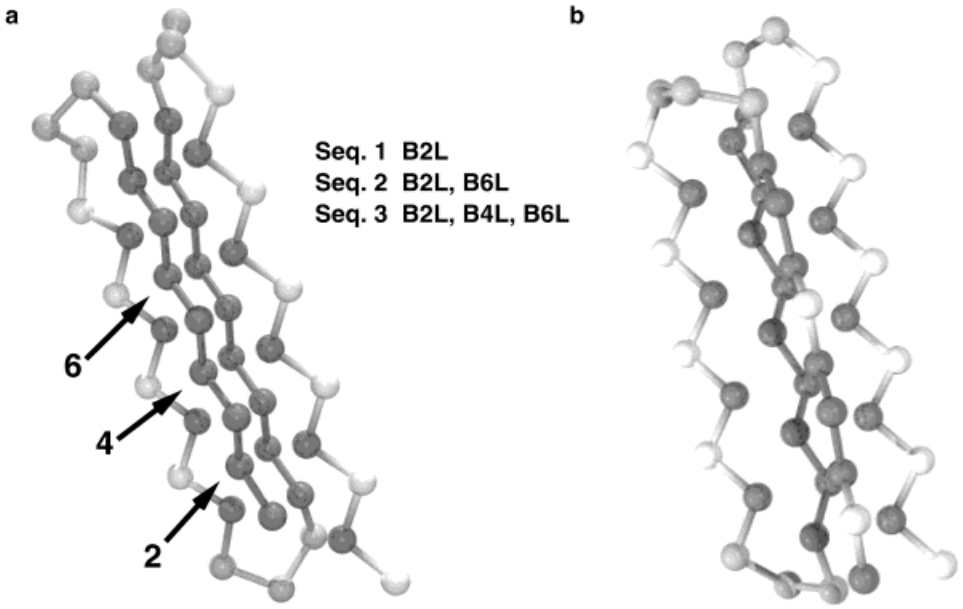

Fig. 1. **a:** Sites for the proposed mutations. The three sequences were modifications of the original Honeycutt and Thirumalal sequence[13] shown here in its native-state structure. Hydrophobic (B) beads are dark, neutral (N) beads are lighter, and hydrophilic (L) beads are the lightest. **b:** Native-state structure for sequence two.

One would expect that in real proteins, sequence diversity and side-chain packing act to make the native-state hydrophobic core significantly favored. In the present work, we took the approach of subtly mutating the hydrophobic core by sequence design to produce a better folding model with a slightly different β-sheet structure, without using a priori knowledge of native contacts or the artificial biases used in previous work. Because we proposed to redesign the hydrophobic core by mutations, we needed a sequence design protocol for proposing beneficial new sequences. We believe the procedure described here is a novel approach to sequence design that is based on threading techniques used in inverse folding,[25] although it was also motivated in part by existing design strategies.[26–28] The newly designed sequences, when compared with the original sequence, were found to possess significantly improved energy landscapes with less traps and long-lived intermediates, lower collapse temperatures, higher folding temperatures, and much faster kinetics.

MATERIALS AND METHODS

This work uses an off-lattice model of a small all-β protein originally proposed by Honeycutt and Thirumalai.[14] The protein chain is modeled as a chain of 46 beads of three flavors: hydrophobic (B), hydrophilic (L), or neutral (N). Attraction between the hydrophobic beads provides the energetic driving force for formation of a strong core, repulsion between the hydrophilic beads, and other beads are used to balance the forces and bias the correct native fold; the neutral beads are treated as relatively floppy residues and signal the turn regions in the sequence. The native-state structure and sequence are shown in Figure 1. We also show in Figure 1 the proposed mutations derived from our sequence design approach described below.

The hamiltonian for the model is[18,23]

$$H = \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} [A(1 + \cos\phi) + B(1 + \cos 3\phi)] + \sum_{i,j \geq i+3} 4\epsilon_H S_1 \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - S_2 \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \quad (1)$$

The bond angles are maintained by a harmonic potential with force constant $k_\theta = 20\epsilon_H/(\text{rad}^2)$ and equilibrium bond angle $\theta_0 = 105°$. The dihedral potential has three minima corresponding to a *trans* state and two *gauche* states. For dihedrals containing two or more neutral residues, $A = 0$ and $B = 0.2\epsilon_H$; for all other dihedrals $A = B = 1.2\epsilon_H$. The nonlocal interactions are given by $S_1 = S_2 = 1$ for BB interactions, $S_1 = 2/3$ and $S_2 = -1$ for LL and LB interactions, and $S_1 = 1$ and $S_2 = 0$ for all interactions involving N residues.

Simulated Annealing and Langevin Simulations.

The global minimum structure for each new sequence was found through simulated annealing. The simulations are performed in reduced units, with the units of mass m, length σ, energy ϵ_H, and k_B all set equal to one; temperature is in units of ϵ_H/k_B. We started with a high-temperature random coil and slowly decreased the

temperature to 0.1. This structure was then steepest-descent quenched to its local minimum energy structure, and the result was saved. The chain was then reheated to 0.65, and the cooling to 0.1 and quench were repeated. This cycle was repeated five times, and the lowest energy structure from these cycles was saved. Ten such simulated annealing runs were performed, for a total search of more than 50 possible global minimum energy structures. The results shown below use this same protocol for finding the global minimum for the original and mutant sequences.

Constant-temperature simulations were performed by using the low-friction limit of Langevin dynamics used in previous studies with the model.[18,19,21,23] The friction coefficient was set to $0.05\tau^{-1}$ and a time step of 0.005τ was used to integrate the equations of motion, where $\tau = \sqrt{m\sigma^2/\epsilon_H}$ is the unit of reduced time. The RATTLE algorithm[29] was used to keep the bond lengths rigid.

Trajectories for histogram analysis were prepared by taking high temperature random coil structures and equilibrating for $1{,}000\tau$ at the desired temperature before taking statistics for $50{,}000\tau$. At most temperatures, five such trajectories were prepared for comprehensive statistics. At lower temperatures, the initial conditions highly influence the sampling, and our strategy was reversed: the equilibration time was stretched to $5{,}000\tau$ (prepared by slow cooling), statistics were only collected for $2{,}500\tau$, with more time between samples to prevent highly correlated sampling, and 60 such trajectories were prepared for each temperature.

Mean first passage times were calculated by taking high-temperature unfolded structures and recording the time that they first folded to the native state at a given temperature.[22] The "dead time" of our calculations, 100τ, results from the initial equilibration phase when the chain is quickly cooled to the appropriate temperature. One could completely eliminate the dead time by rescaling the velocities of the random coil at $t = 0$ to give a structure with the proper temperature, but a structure prepared in this manner would not necessarily be a member of the proper canonical ensemble at that temperature. Our quick cooling approach is an attempt to produce a more representative unfolded chain for a given temperature.

The Multiple Histogram Method

The multiple multidimensional histogram method was used to probe the free energy landscape of the original and mutant sequences.[21,23,30] The relevant equations and procedure for determining the density of states have been well described in the past.[30,31] Three-dimensional histograms over energy, radius of gyration R_g, and native-state similarity χ were collected as described above. The native-state similarity measure, χ, was implemented as described in previous work[18]:

$$\chi = 1 - \frac{2}{2 + N(N-3)} \sum_{i,j \geq i=2}^{N} \theta(\epsilon - |r_{ij} - r^N_{ij}|) \qquad (2)$$

where the double sum is over beads on the chain, r_{ij} and r^N_{ij} are the distances between beads i and j in the state for comparison and the native state, respectively, θ is the Heaviside step function, and $\epsilon = 0.2$ to account for small fluctuations away from the native-state structure. χ ranges from values of 0.0, corresponding to structures identical to the native state, to values of $\approx$0.9 for random coil structures.

One advantage of histogram methods is the ability to extrapolate thermodynamic observables to temperatures where well-converged values are not easily accessible on the timescale of ordinary simulations. However, contingent on this use is that the underlying energy landscape is not very glasslike. If low-lying intermediates and traps populate the free energy landscape at low temperatures, then histogram techniques alone are not very useful. Sampling at low temperatures will not give properly weighted histograms because the states sampled will sensitively depend on initial conditions. Sampling at higher temperatures cannot overcome this problem because the fine features of the underlying glass landscape will be rarely sampled at higher energies, and these details are essential in determining the true low temperature thermodynamics. We found these issues to be most problematic for determining the low temperature behavior of the original sequence, similar to the conclusion of Nymeyer et al.[22] Recent techniques for more accurate sampling of glasslike landscapes such as multicanonical and generalized ensemble methods,[32–34] entropic sampling,[35] umbrella sampling of the potential energy,[36] or jump-walking[37] can be combined with the histogram method to overcome these problems.

RESULTS

Sequence Design Procedure

With this library of misfolded structures and the native state structure in hand, it was a simple matter to thread new sequences through the structures and examine the resulting distribution of energies. Good sequences would be expected to a have a significant separation in energy between the native and next lowest excited state[38] and would be expected to possess a large separation between the average misfolded energy and the energy of the native state.[13]

The three sequences proposed in this article were selected by setting which type of mutation was desired (one, two, or three beads) and maximizing the energy gap between the native and next low-lying state. The proposed mutations were made to the hydrophobic core, replacing one or more hydrophobic (B) beads with hydrophilic (L) beads. Our sequence design procedure proposed these mutations by threading all possible single, double, and triple $B \rightarrow L$ mutant sequences through the library of collapsed states and choosing sequences that possessed a highly favorable energy gap between the native and misfolded states. Figure 1 summarizes the locations of the mutation sites and the resulting sequences.

Several improvements to this design scheme can be readily suggested. An obvious way to improve the thread-

ing approach would be to increase the size of the library of compact states. Although requiring dissimilar states allows a smaller number of conformations to be representative of the space of compact structures, a hundred structures is probably still too few to capture the full possibility of misfolded structures. Furthermore, the structures in our current library were not quenched to a local minimum, and we would expect that doing so would also make each structure more representative of a potential trap. One could also propose a recursive scheme for designing better and better sequences, in which each new sequence is used to generate a new library of representative misfolded structures, and the sequence design strategy is repeated on this new library to generate a new sequence. Our strategy did fail on our first attempt to design a triple-mutant sequence; the proposed sequence possessed far too many traps in simulated annealing runs. Sequence three was the second best sequence from the design trials for triple mutants, but it performed better as a folding sequence.

In spite of these suggested improvements, the results presented next show that our proposed design strategy works well in the present case. An interesting extension of this approach would be to "thread" new energy functions through the structure library in a search for a hamiltonian with a better energy landscape.[23,39] For example, we could use this approach to search for energy terms modeling solvation that could potentially improve protein folding models.[40]

Finding Global Minimum Structures Using Simulated Annealing

The global minimum structures for each sequence were found by using a simulated annealing protocol described above. Of course, we had no guarantee that new sequences would have the same global minimum structure, but the hope was that subtle mutations would not affect the overall β-sheet structure dramatically, and this assumption has proven true for the mutant sequences described here. Figure 1 compares the global minimum structures of the original sequence to that of mutant sequence two. Although the structures are different—the new structure has in common only 22 of 53 native contacts present in the original structure—the overall topology of the β-sheet fold is identical.

In fact, the native-state structure of the mutant sequences is more structurally similar to the low-lying excited state noted above.[20] Introduction of a single mutation lowers the excited state energy below the native state and increases the gap between these states to 9.1 ϵ_H. The replacement of a single *B* bead by an *L* forces only one alignment of the chain to be favored, and the result is an excellent energy separation between these two competing states. In effect, the modification of the hydrophobic core is simulating side-chain packing, by forcing the hydrophobic strands to be in register in a very specific alignment. It is important to note that the native-state structures of all three mutant sequences share a similarity measure of $\chi <$ 0.2, whereas all three structures are dissimilar by $\chi \approx 0.6$ to the original native-state structure.

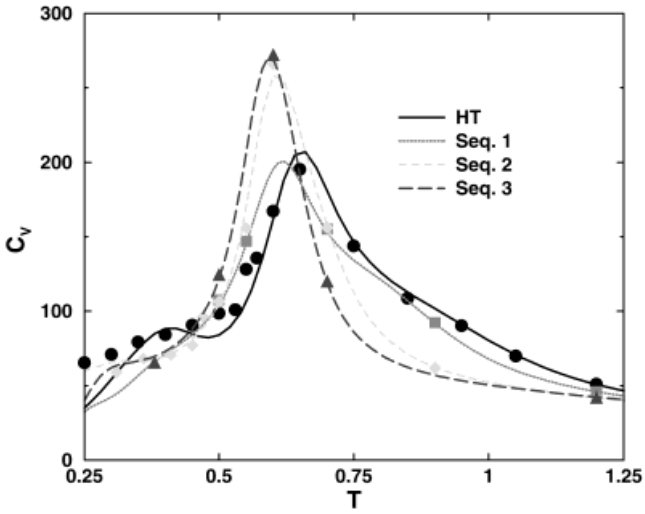

Fig. 2. Heat capacity versus temperature for the original (HT) and mutant sequences. The points are from molecular dynamics simulations at those temperatures. At low temperatures we can see artifacts of inadequate sampling (see Materials and Methods).

The original sequence possesses a very rough energy landscape with many low-lying minima with little energy separation between them. This is manifested in our simulated annealing runs by the observation that the global minimum structure is only found 8% of the time for the original sequence. For each new sequence, simulated annealing runs were conducted to find the new global minimum structure. In contrast, the mutant sequences found their respective global minimum 36%, 57%, and 46% of the time, respectively. This is already indicative that the mutations were beneficial in removing some of the roughness of the underlying landscape and better biasing the native-state structure. Although it might be possible to find a simulated annealing protocol that found the global minimum more often for the original sequence, we would expect that these same results would hold true; that is, the global minimum structure is found much more reliably in the mutant sequences.

Thermodynamics

Using trajectories from Langevin dynamics[18] and the multiple histogram technique[30,31,41], we can characterize the change in folding thermodynamics for the new sequences. The nature of the collapse transition can be determined by examining the heat capacity versus temperature curves for the original and mutant sequences (Fig. 2). We see that the effect of even a single mutation is to shift the collapse temperature (T_θ), the temperature of the peak in $C_V(T)$, to lower temperatures. Second and third mutations in the core do not significantly shift the collapse temperature to lower temperatures; instead, they increase the sharpness of the peak in specific heat, indicative of a more cooperative, two-state transition. This further

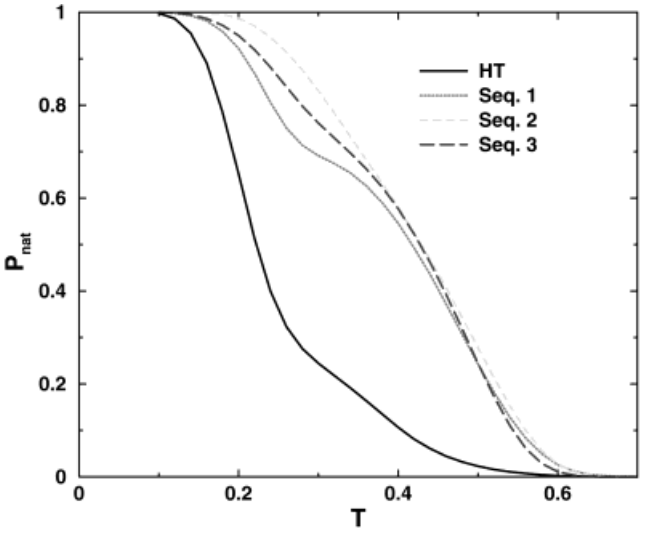

Fig. 3. Native-state population versus temperature for the original and mutant sequences.

strengthens the analogy to the effect of side-chain packing, in that it seems that the core mutations are forcing more cooperativity in the collapse process and favoring more discrete-state kinetics.

Folding to the native state can be best characterized by calculating the native-state population versus temperature. The native state population at a temperature T is defined by

$$P_{nat}(T) = \frac{\sum_{E,\chi<\chi_{nat}} \Omega(E,\chi)e^{-E/T}}{\sum_{E,\chi} \Omega(E,\chi)e^{-E/T}} \quad (3)$$

where $\Omega(E,\chi)$ is the density of states as a function of potential energy E and native state similarity χ, χ_{nat} is the value of χ that defines the native-state basin of attraction,[21] and the Boltzmann constant, k_B is set equal to one. Figure 3 shows the result of this calculation for the original sequence and the three mutant sequences. The values of χ that define the native-state basin of attraction can be determined from a plot of free energy versus χ and identifying the χ value of the transition state.[21] For all of the mutant sequences the folding temperature (T_f), the temperature at which $P_{nat}(T) = 0.5$, is seen to have noticeably increased. The folding curve for sequence one looks possibly three state, but the folding temperatures for all three mutants converge to a value of $T_f \approx 0.4$. Our curve for the original sequence is noticeably different from that plotted in Guo and Brooks[21] and shows the difficulty of using histogram techniques on very glassy energy landscapes.

The folding and collapse transitions are summarized in Table I. Also shown in the table is Thirumalai's σ parameter[42–45] defined as

$$\sigma = \frac{T_\theta - T_f}{T_\theta}, \quad (4)$$

TABLE I. Collapse and Folding Temperatures for the Original and Mutant Sequences†

Sequence	T_θ	T_f	σ
HT	0.65	0.22	0.66
1	0.61	0.42	0.31
2	0.61	0.43	0.30
3	0.59	0.43	0.27

†Also shown is $\sigma = (T_\theta - T_f)/T_\theta$. Temperatures are in units of ϵ_H/k_B.

which has been shown to correlate well with faster folding kinetics. The lower values of σ found for the mutant sequences coincide with values based on experiments on three-state proteins,[44] indicating that the mutants might proceed through this type of folding pathway. We will show below that the lower values are also consistent with much faster folding kinetics and a delayed glass transition. The value of σ for the original sequence is significantly higher than that observed in any experiments on small proteins.[44] The exact value of σ for the original sequence depends on the folding temperature that, as noted before, is very difficult to precisely calculate. Even with the previously reported[21] higher T_f value of 0.35, the value of σ for the original sequence is higher than that expected for small proteins.

An advantage of multidimensional histogram methods is the ability to project the free energy landscape onto multiple order parameters. For this model it has proven useful to examine the free energy surface as a function of both the collapse order parameter R_g, and the folding order parameter χ.[21,23] Note that our definition of χ (see above) follows earlier work[18,21] in that a value of $\chi = 0$ corresponds to the native state and $\chi \approx 0.9$ corresponds to a random coil; this is one minus the definition used in Shea et al.[23]

The original surface possessed a very strong "L" shape, indicating that folding had to occur by the chain first collapsing to a non-native state and then suitably rearranging. Faster folding, with less traps, can be achieved if the surface is pulled more toward the diagonal, i.e., allowing acquisition of native contacts as the chain is collapsing.[23]

We can see from Figure 4 that the mutant sequences do indeed improve the folding free energy surface by allowing more nativelike states when partially collapsed. A similar result had been achieved before by specifically biasing formation of native-state contacts.[23] Here we demonstrate that the properties of the sequence alone can modify the folding free energy surface in a beneficial manner. Comparing the very bottom of the surfaces, we can also see that the most compact, non-native states present in the original sequence are entirely missing in mutant sequence two. Similar results hold true for the remaining redesigned sequences.

Kinetics

Having established that the mutant sequences possess a better underlying free energy landscape in quantities such as collapse and folding temperatures, it was of great

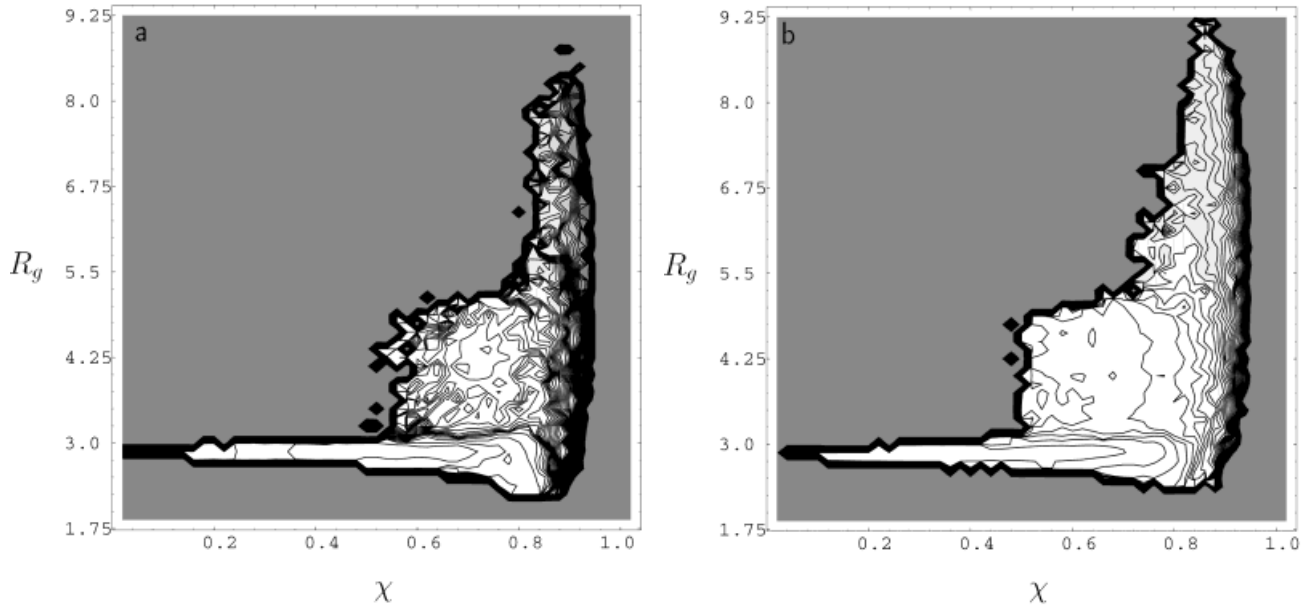

Fig. 4. **a:** $R_g - \chi$ free energy surface at $T = 0.3$ for the original sequence. **b:** $R_g - \chi$ free energy surface at $T = 0.4$ for sequence two. The contour lines are spaced $3k_BT$ apart.

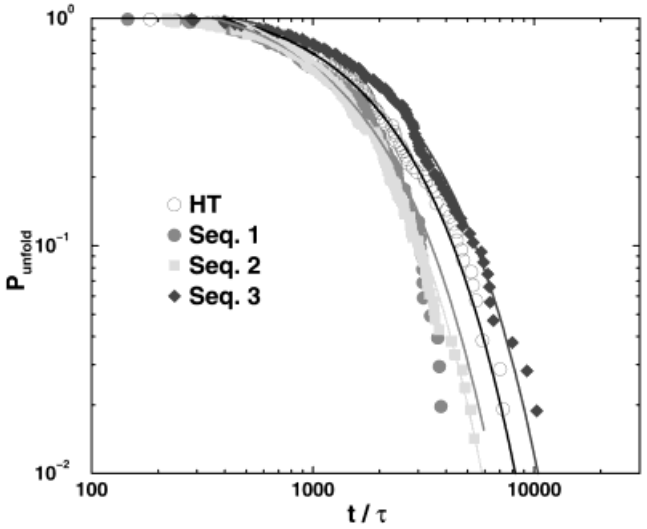

Fig. 5. Unfolded population versus time at $T = 0.6$ for original and mutant sequences. The curves are single exponential fits to the data points.

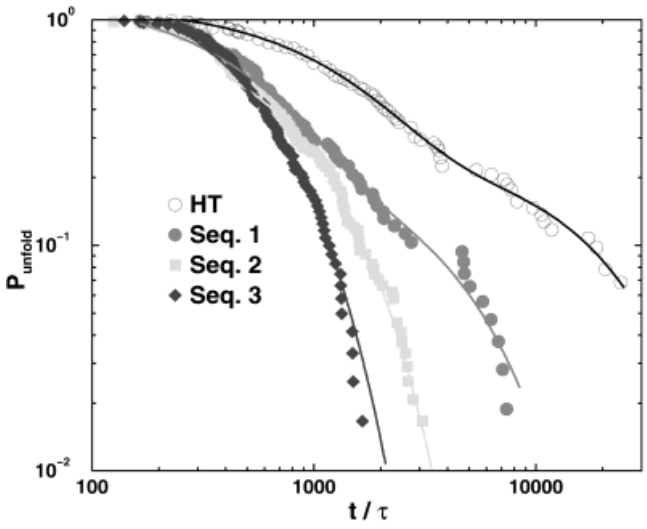

Fig. 6. Unfolded population versus time at $T = 0.5$ for original and mutant sequences. The curves for the original sequence and sequence one are biexponential fits to the data points. The remaining curves are single exponential fits.

interest to determine the resulting change in folding and unfolding kinetics.

Folding kinetics were examined by tabulating folding first-passage times as a function of temperature. Figure 5 shows the relatively high-temperature folding kinetics of the original and mutant sequences. $T = 0.6$ is just below or around the collapse transition for these sequences, and we can see that the resulting kinetics are extremely similar for all of the sequences. Folding to the native state is relatively slow at this temperature because the native state is unstable, and the barrier for reaching the native state is expected to be consequently higher. At this temperature, the folding scenario is mostly two-state, as can be seen by the adequacy of single exponential fits to the folding kinetics.

The expected differences between the sequences become observable as we proceed to study folding at lower temperatures (Fig. 6). Now evidence of biexponential or multiexponential kinetics is present in the original sequence and, to some extent, sequence one. This is to be expected for landscapes possessing many long-lived traps and/or intermediates. However, sequences two and three have significantly destabilized traps to the point that the kinetics are

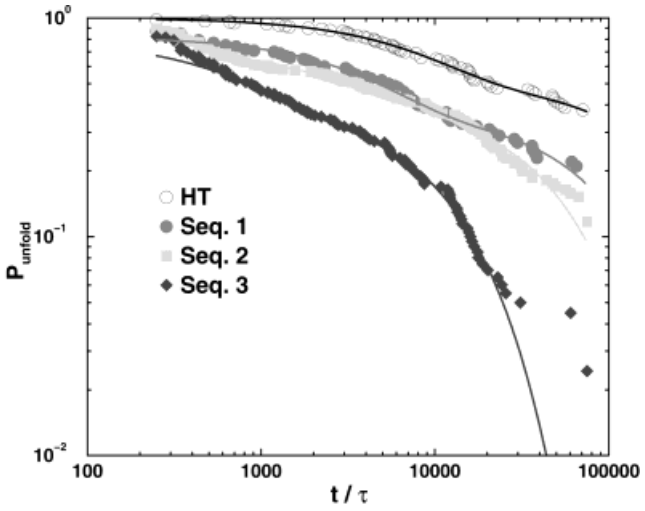

Fig. 7. Unfolded population versus time at $T = 0.35$ for the original and mutant sequences. The curves are biexponential fits to the data points.

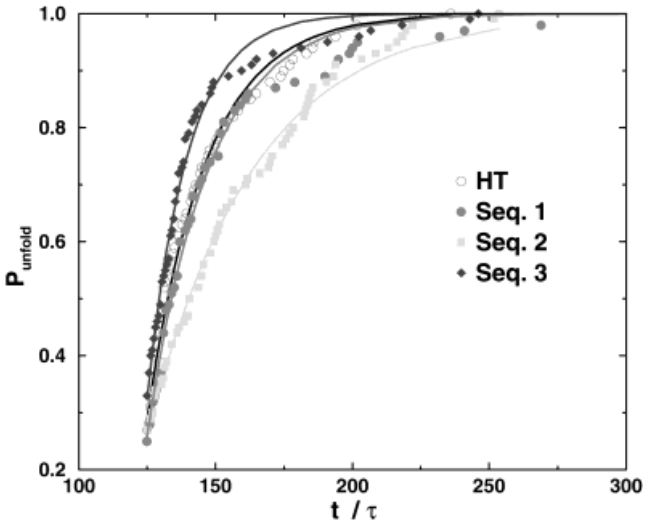

Fig. 8. Unfolded population versus time for unfolding of the original and mutant sequences at $T = 0.4$. The curves are single exponential fits (see Discussion).

still single exponential and significantly faster than the original sequences.

As we proceed to lower temperatures, folding becomes increasingly slower for each sequence. Figure 7 shows the low-temperature kinetics at $T = 0.35$, below the folding temperature for the mutant sequences, but still above the folding temperature for the original sequence. The folding runs were conducted for a longer time ($1.5 \times 10^6\tau$) at this temperature for more accurate data. On the observed timescale, the data still fit a biexponential fairly well. We might expect lower temperature kinetics with many traps to be better fit with a stretched exponential[46] ($e^{-(t/\tau)^\beta}$), but this form was not found to fit any better. The biexponential kinetics are characterized by a relatively fast phase (800–8,500τ) and a slow phase (12,000–240,000τ), with the original sequence possessing the slowest time constants and sequence three, the fastest.

A dramatic indication of the improvement in kinetics from mutating the core is that sequence three found the native state in 97% of the kinetic trials, whereas the original sequence only found the native state 60% of the time. Surprisingly, sequences one and two fold with similar looking kinetics at $T = 0.35$, although the time constant for the slow folding population is nearly twice as long for sequence one. Deviations from biexponential behavior can be seen in sequences two and three at long times, possibly indicating additional traps.

The ratio of the folding temperature to the glass temperature is of interest for comparing simplified models to experiments.[24] Although in the present case, it is difficult to quantify the glass temperature, we can make conclusions about whether the folding temperature is above the glass temperature for these sequences—a characteristic of a good folder.[47] Below the glass temperature, we would expect the kinetics to follow a power law behavior.[22] Such a behavior would appear as a straight line on a log-log plot. Although the $T = 0.35$ kinetics of the original sequence and sequence one adequately fit biexponentials on the timescales examined here, we can also see that power law behavior has begun to set in. However, at this low temperature, sequences two and three still show curved kinetic behavior, characteristic of exponential kinetics. On the basis of these observations, we can conclude that the original sequence possesses a folding temperature below the glass temperature,[22] and sequences two and three possess glass temperatures below their folding temperatures. Sequence one probably possesses a glass temperature below its folding temperature, because the kinetics still have some curvature at $T = 0.35$ and the folding temperature is $T_f = 0.42$, but the folding temperature is much closer to the glass temperature in this case.

A side effect of our sequence design strategy for destabilizing traps is that, to some extent, the native-state hydrophobic core is destabilized as well. We can quantity the native-state stability by monitoring unfolding from simulations started in the native state structure (Fig. 8). The unfolding kinetics at $T = 0.4$ are well fit by a single exponential of the form

$$P_{unfold}(T) = \exp\left[-\frac{t - t_0}{\tau_u}\right] \tag{5}$$

where t_0 is the time lag necessary for taking the zero-temperature native state and bringing it to a $\langle\chi\rangle_{T=0.4}$ structure, and τ_u is the time constant for unfolding. Some deviation from this form can be seen in sequences one and three, which could be taken as evidence of two or more distinct basins in the native-state basin of attraction. This plot shows that sequence two is the most stable at this temperature, and sequence three is least stable, although none of the sequences are particularly stable, considering how relatively easy it is to leave the native state.

Inherent in measuring mean first passage times is the definition of the native state. Because a previous study[21] had established $\chi \leq 0.3$ as a sufficient definition, this definition was used for all of the kinetics plots shown in Figures 5–7. A choice of $\chi \leq 0.4$ can be justified for the mutant sequences, because this is the location of the maximum in free energy versus χ for those sequences (data not shown). Such a choice would make the observed kinetics even faster, although we would not expect much change because once the chain has passed the $\chi = 0.4$ barrier, it should be relatively quick to reach $\chi = 0.3$. The folding kinetics were reevaluated for sequence two with the looser $\chi < 0.4$ choice, and we found that at higher temperatures this is indeed true; the folding times for reaching $\chi = 0.3$ are only slightly longer than the times to reach $\chi = 0.4$. At lower temperatures, the folding times become much slower for the $\chi \leq 0.3$ criterion. This indicates the possible presence of intermediates not visible in the one-dimensional projection of free energy onto χ. However, for fairness of comparison we used the slower $\chi \leq 0.3$ choice for all sequences in Figures 5–7.

DISCUSSION

We earlier compared the effect of our mutations to the effect of side-chain packing. Simulated annealing showed that low-lying traps were destabilized relative to the global minimum, an effect that would be expected if side-chain packing favoring the native hydrophobic core had been included. Evidence for more cooperativity in the collapse and folding transitions is also in agreement with this observation.[44] However, we can also make sense of our more cooperative transitions by the observation that our mutations have effectively made the chain energetics more repulsive, and a collapse transition concomitant with formation of the native state has been observed in many lattice models when interactions are made more repulsive.[40,46,48–51] The analogy to side-chain packing is also weakened when we examine the unfolding kinetics. One would expect that true side-chain packing would also provide a heightened kinetic barrier to unfolding of the native state. Although a specific native state is more highly favored in our sequences, the accompanying barrier to leaving the native state does not look appreciably higher.

The effects of the core mutations can also be understood in terms of a recent classification scheme for folding scenarios developed by Wolynes and colleagues.[5,11,52] The original sequence possesses many long-lived misfolded intermediates, and these dominate the folding kinetics at temperatures below T_θ. Such a folding scenario would be termed type IIB.[5,11] In contrast, the kinetics exhibited by the mutant sequences, particularity sequences two and three, are more likely a type IIA or possibly type I folding scenario—one in which the kinetics are faster and glass-like dynamics are only encountered below the folding temperature. Because the low-temperature details of systems that exhibit glass behavior are extremely sequence-dependent,[11,52], i.e., not self-averaging, we would *a priori* expect from theoretical considerations alone that mutations can significantly change the low-temperature aspects of a particular protein model.

Another issue encountered in this work is the question of a suitable definition of the native state. A natural definition is one based on structural similarity to the global minimum structure. However, there are various ways to implement this, including the χ definition used here,[18] considering nearly all pairwise distances, or a definition based on native contacts,[22] considering only pair distances of native-state local contacts. The χ definition is more stringent than a definition based solely on local contacts, and some of the quantitative aspects of the thermodynamics and kinetics of the folding transition found in this work would change if a looser native-state criterion was used. However, we used the χ criterion for all of the sequences studied here, so our results can be quantitatively compared within this work. We would certainly expect that any other reasonable definition of the native state would find similar improvements in the thermodynamics and folding kinetics for the mutant sequences versus the original sequence.

Experimental Comparision

The high-temperature ($T = 0.6$) folding of these sequences is best characterized with a single exponential and is indicative of a single rate-limiting barrier in the folding pathway. At lower temperatures the folding becomes multiexponential, with a fraction of the chains folding quickly to the native state and a fraction collapsing to long-lived misfolded structures and folding much more slowly. Such multiexponential kinetics has been observed before in many fast-folding experiments on small proteins.[53–59] For instance, similar to the present model, cytochrome c has been observed to follow a multiple pathway scenario with either an initial collapse to the native state or collapse to a compact intermediate followed by folding to the native state.[55,56]

The observation from this work that a single mutation in the hydrophobic core can deeply influence the thermodynamics and kinetics of folding has also been noted several times before in an experimental context. Destabilizing mutations in the hydrophobic core of a Trp-containing ubiquitin are known to change the kinetics from three-state to two-state, whereby the folding intermediate found in the wild-type folding pathway is destabilized to the point that it no longer noticeably accumulates.[60] Single mutations in the core of a 98 amino acid β-sheet protein were also found to significantly affect the stability of a collapsed folding intermediate, with a consequent change in the speed of folding.[61] Similarly, a double Gly→ Ala mutation in the 80 amino acid fragment of λ repressor changes the structure of the transition state and dramatically affects the folding pathway.[62] In contrast, folding studies by Schmid and co-workers[63] found that sequence differences in small all-β cold-shock proteins do not lead to an appreciably different kinetic mechanism; however, the residues differing between proteins were all surface residues. They note that changing the interior residues, as we

have done here, would likely be influential to the mechanism of folding.[63]

An interesting link between experiment and this work is the de novo design study of a four-helix bundle by Handel et al.[64] They originally designed a sequence to fold to a four-helix bundle stabilized primarily by a large hydrophobic core of leucines. This original sequence formed a collapsed state stabilized by hydrophobic forces but lacked the specific interactions necessary for forming a true native state.[64] By adding two three-His Zn^{2+}-binding sites into the original sequence they found that they could bias the sequence to fold to a unique native state structure in the presence of Zn^{2+} as characterized by NMR and ANS binding. The introduction of specific interactions into a degenerate hydrophobic core was sufficient to produce behavior much more resembling biological proteins.

CONCLUSION

The essence of any design strategy for protein sequences is to not only ensure stabilization of the native state but to also destabilize non-native structures.[13,26,65] Often this procedure is attempted in a mean-field approach, where misfolded structures are not known previous to the design step, and the average misfolded energy is estimated. Here we have taken a simplified model of a protein with many known traps and long-lived intermediates and used this specific structural information to constructively modify the protein sequence.

That such a technique can effectively work is evidenced most by the simplest mutation, a single substitution of a hydrophilic bead for a hydrophobic bead. Sequence one differs at a single position from the original sequence, yet simulated annealing runs find the global minimum structure more reliably, indicating a much smoother energy landscape. The collapse temperature is lowered because of the introduced repulsion, and this coincides nicely with the raised folding temperature due to the destabilization of non-native misfolds. The net effect is that a single mutation changes the folding scenario from one dominated by many low-energy traps to a more simultaneous collapse and folding process with lower barriers for rearrangement of collapsed structures to the native state. Although the high symmetry of the original sequence codes well for the desired native β-sheet structure, the cost of this overengineered symmetry is a highly degenerate hydrophobic core. Breaking the symmetry with the single mutation destabilizes many non-native cores and adds the frustration necessary for successful discrimination of the native structure.[13]

A second mutation in the core further improves the thermodynamics and kinetics to the point that we can confidently assert that, unlike the original sequence,[22] the glass temperature has been lowered below the folding temperature. The higher folding temperature of this sequence also results in the highest native-state stability at $T = 0.4$ of the sequences examined here. Because the kinetics have been appreciably improved for this sequence, with a successful raising of the folding temperature, we would promote this sequence as being most proteinlike and therefore interesting for future protein folding studies that might investigate folding mechanisms or general characteristics of the folding funnel.

Adding a third mutation to the core continues to improve the kinetics, pushing the glass temperature further below the folding temperature. However, in this case, the cost of fast-folding kinetics is some loss in stability of the native state. Experimentally, it has been observed that many fast-folding small proteins also sacrifice stability for this property.[66–68] Some recent experiments have even shown no clear correlation between folding speed and native stability.[69,70] At present, it appears that the positive correlation between higher native state stability and faster folding speed noted in some earlier lattice model studies[38] is not necessarily a general feature for all proteins.

Modeling attempts in the last decade have made it clear that correctly modeling the free energy landscape for proteins is a subtle task. Simplified models concentrating on physical interactions often encounter glassy behavior at temperatures before folding,[22,40,47] but models that avoid this behavior do so by explicitly biasing the native state to be inherently more favorable. We have shown that problematic models based on physical interactions can be rehabilitated with a threading approach to sequence design. The more proteinlike mutant sequences reported here should prove useful for future protein folding studies.

REFERENCES

1. Go N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
2. Dill KA, Bromberg S, Yue K, et al. Principles of protein folding-a perspective from simple exact models. Protein Sci 1995;4:561–602.
3. Shakhnovich EI. Theoretical studies of protein-folding thermodynamics and kinetics. Curr Opin Struct Biol 1997;7:29–40.
4. Pande VS, Grosberg AY, Tanaka T, Rokhsar DS. Pathways for protein folding: is a new view needed? Curr Opin Struct Biol 1998;8:68–79.
5. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.
6. Brooks CL. Simulations of protein folding and unfolding. Curr Opin Struct Biol 1998;8:222–226.
7. Dobson CM, Sali A, Karplus M. Protein folding: a perspective from theory and experiment. Angew Chem Int Ed Engl 1998;37:868–893.
8. Harrison PM, Chan HS, Prusiner SB, Cohen FE. Thermodynamics of model prions and its implications for the problem of prion protein folding. J Mol Biol 1999;286:593–606.
9. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
10. Leopold PE, Montal M, Onuchic JN. Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci USA 1992;89:8721–8725.
11. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195.
12. Dill KA, Chan HS. From Levinthal to pathways and funnels. Nature Struct Biol 1997;4:10–19.
13. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84:7524–7528.
14. Honeycutt JD, Thirumalai D. Metastability of the folded states of globular proteins. Proc Natl Acad Sci USA 1990;87:3526–3529.
15. Honeycutt JD, Thirumalai D. The nature of folded states of globular proteins. Biopolymers 1992;32:695–709.

16. Guo Z, Thirumalai D, Honeycutt JD. Folding kinetics of proteins: a model study. J Chem Phys 1992;97:525–535.
17. Thirumalai D, Guo Z. Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labeling experiments. Biopolymers 1994;35:137–140.
18. Guo Z, Thirumalai D. Kinetics of protein folding: nucleation mechanism, time scales, and pathways. Biopolymers 1994;36:83–102.
19. Guo Z, Thirumalai D. The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. Fold Des 1997;2:377–391.
20. Berry RS, Elmaci N, Rose JP, Vekhter B. Linking topography of its potential surface with the dynamics of folding of a protein model. Proc Natl Acad Sci USA 1997;94:9520–9524.
21. Guo Z, Brooks CL. Thermodynamics of protein folding: a statistical mechanical study of a small all-β protein. Biopolymers 1997;42: 745–757.
22. Nymeyer H, Garcia AE, Onuchic JN. Folding funnels and frustration in off-lattice minimalist protein landscapes. Proc Natl Acad Sci USA 1998;95:5921–5928.
23. Shea JE, Nochomovitz YD, Guo Z, Brooks CL. Exploring the space of protein folding hamiltonians: the balance of forces in a minimalist β-barrel model. J Chem Phys 1998;109:2895–2903.
24. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. Toward an outline of the topography of a realistic protein-folding funnel. Proc Natl Acad Sci USA 1995;92:3626–3630.
25. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
26. Abkevich VI, Gutin AM, Shakhnovich EI. Improved design of stable and fast-folding model proteins. Fold Des 1996;1:221–230.
27. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 1993;90:7195–7199.
28. Hao MH, Scheraga HA. Optimizing potential functions for protein folding. J Phys Chem 1996;100:14540–14548.
29. Andersen HC. Rattle: a "velocity" version of the Shake algorithm for molecular dynamics calculations. J Comp Phys 1983;52:24–34.
30. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. Multidimensional free-energy calculations using the weighted histogram analysis method. J Comp Chem 1995;16:1339–1350.
31. Ferrenberg AM, Swendsen RH. Optimized Monte Carlo data analysis. Phys Rev Lett 1989;63:1195–1198.
32. Berg BA, Neuhaus T. Multicanonical algorithms for first order phase transitions. Phys Lett 1991;267:249–253.
33. Nakajima N, Nakamura H, Kidera A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. J Phys Chem B 1997;101:817–824.
34. Hansmann UHE, Okamoto Y, Onuchic JN. The folding funnel landscape for the peptide Met-enkephalin. Proteins 1999;34:472–483.
35. Hao MH, Scheraga HA. Monte Carlo simulation of a first-order transition for protein folding. J Phys Chem 1994;98:4940–4948.
36. Bartels C, Karplus M. Probability distributions for complex systems: adaptive umbrella sampling of the potential energy. J Phys Chem B 1998;102:865–880.
37. Tsai CJ, Jordan KD. Use of the histogram and jump-walking methods for overcoming slow barrier crossing behavior in Monte-Carlo simulations—applications to the phase transitions in the Ar_{13} and $(H_2O)_8$ clusters. J Chem Phys 1993;99:6957–6970.
38. Sali A, Shakhnovich EI, Karplus M. How does a protein fold? Nature 1994;369:248–251.
39. Park B, Levitt M. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
40. Sorenson JM, Head-Gordon T. The importance of hydration for the kinetics and thermodynamics of protein folding: simplified lattice models. Fold Des 1998;3:523–534.
41. Ferrenberg AM, Swendsen RH. New Monte Carlo technique for studying phase transitions. Phys Rev Lett 1988;61:2635–2638.
42. Klimov DK, Thirumalai D. Criterion that determines the foldability of proteins. Phys Rev Lett 1996;76:4070–4073.
43. Veitshans T, Klimov D, Thirumalai D. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. Fold Des 1996;2:1–22.
44. Klimov DK, Thirumalai D. Cooperativity in protein folding: from lattice models with sidechains to real proteins. Fold Des 1998;3: 127–139.
45. Klimov DK, Thirumalai D. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. J Chem Phys 1998;109:4119–4125.
46. Socci ND, Onuchic JN, Wolynes PG. Protein folding mechanisms and the multidimensional folding funnel. Proteins 1998;32:136–158.
47. Socci ND, Onuchic JN. Folding kinetics of proteinlike heteropolymers. J Chem Phys 1994;101:1519–1528.
48. Gutin AM, Abkevich VI, Shakhnovich EI. Is burst hydrophobic collapse necessary for protein folding? Biochemistry 1995;34:3066–3076.
49. Socci ND, Onuchic JN. Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique. J Chem Phys 1995;103:4732–4744.
50. Camacho CJ, Thirumalai D. Denaturants can accelerate folding rates in a class of globular proteins. Protein Sci 1996;5:1826–1832.
51. Chan HS, Dill KA. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. Proteins 1998;30:2–33.
52. Wolynes PG, Luthey-Schulten Z, Onuchic JN. Fast-folding experiments and the topography of protein folding energy landscapes. Chem Biol 1996;3:425–432.
53. Udgaonkar JB, Baldwin RL. Early folding intermediate of ribonuclease A. Proc Natl Acad Sci USA 1990;87:8197–8201.
54. Briggs MS, Roder H. Early hydrogen-bonding events in the folding transition of ubiquitin. Proc Natl Acad Sci USA 1992;89:2017–2021.
55. Sosnick TR, Mayne L, Hiller R, Englander SW. The barriers in protein folding. Nature Struct Biol 1994;1:149–156.
56. Shastry MCR, Roder H. Evidence for barrier-limited protein folding kinetics on the microsecond time scale. Nature Struct Biol 1998;5:385–392.
57. Morgan CJ, Miranker A, Dobson CM. Characterization of collapsed states in the early stages of the refolding of hen lysozyme. Biochemistry 1998;37:8473–8480.
58. Eaton WA, Muñoz V, Thompson PA, Henry ER, Hofrichter J. Kinetics and dynamics of loops, α-helices, β-hairpins, and fast-folding proteins. Acct Chem Res 1998;31:745–753.
59. Kulkarni SK, Ashcroft AE, Carey M, Masselos D, Robinson CV, Radford SE. A near-native state on the slow refolding pathway of hen lysozyme. Protein Sci 1999;8:35–44.
60. Khorasanizadeh S, Peters ID, Roder H. Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. Nature Struct Biol 1996;3:193–205.
61. Lorch M, Mason JM, Clarke AR, Parker MJ. Effects of core mutations on the folding of a β-sheet protein: implications for backbone organization in the I-state. Biochemistry 1999;38:1377–1385.
62. Burton RE, Huang GS, Daugherty MA, Fullbright PW, Oas TG. Microsecond protein folding through a compact transition state. J Mol Biol 1996;263:311–322.
63. Perl D, Welker C, Schindler T, et al. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. Nature Struct Biol 1998;5:229–235.
64. Handel TM, Williams SA, DeGrado WF. Metal ion-dependent modulation of the dynamics of a designed protein. Science 1993; 261:879–885.
65. Yue K, Dill KA. Inverse protein folding problem: designing polymer sequences. Proc Natl Acad Sci USA 1992;89:4163–4167.
66. Schindler T, Schmid FX. Thermodynamic properties of an extremely rapid protein folding reaction. Biochemistry 1996;35: 16833–16842.
67. Tang KS, Guralnick BJ, Wang WK, Fersht AR, Itzhaki LS. Stability and folding of the tumour suppressor protein pl6. J Mol Biol 1999;285:1869–1886.
68. Ladurner AG, Itzhaki LS, Fersht AR. Strain in the folding nucleus of chymotrypsin inhibitor 2. Fold Des 1997;2:363–368.
69. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
70. Kim DE, Gu H, Baker D. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. Proc Natl Acad Sci USA 1998;95:4982–4986.